Project proposal: Analyze on the 2018 Kaggle Machine Learning and Data Science Survey database

I want my profession to be a data scientist in near future so some questions about the current state of data science community comes to the mind. The best large available dataset I could find to start analyzing was “The 2018 Kaggle Machine Learning and Data Science Survey”. The challenge is to deeply explore (through data) the impact, priorities, or concerns of a specific group of data science and machine learning practitioners. That group can be defined in the macro (for example: anyone who does most of their coding in Python) or the micro (for example: female data science students studying machine learning in masters programs). This is an opportunity to be creative and tell the story of a community you identify with or are passionate about!

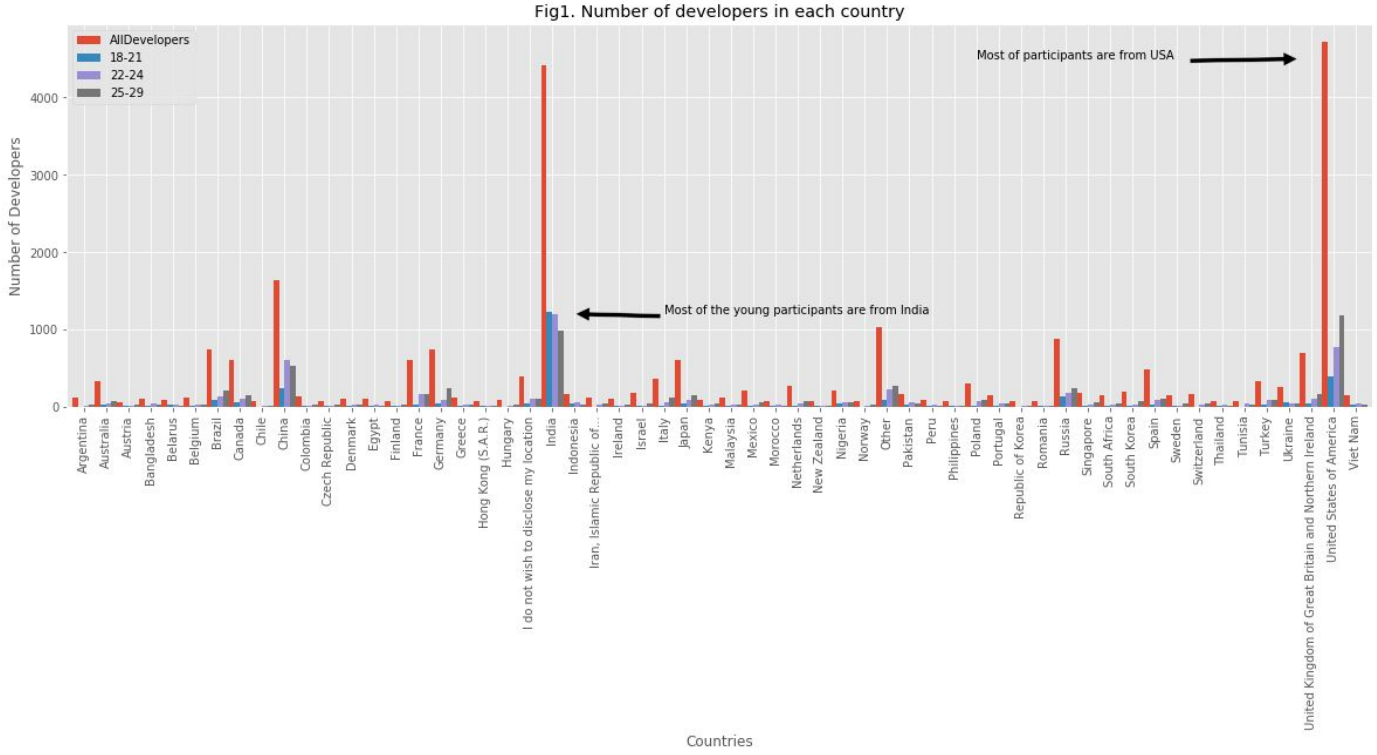
The 2018 Kaggle Machine Learning and Data Science Survey finished with 23,859 responses. Questions and responses are available in multipChoiceResponses.csv.

50 Questions were asked, here are some examples:

- Q1 : What is your gender?
- Q2 : What is your age (# years)?
- Q3 : In which country do you currently reside?
- What is your age (# years)?
- Q4 : What is the highest level of formal education that you have attained or plan to attain within the next 2 years?
- Q46 : Approximately what percent of your data projects involve exploring model insights?

To observe the distribution of participants, I plot the number of participants from different country. Also to see which countries have the youngest developers, also plot the participant in age range of [18-21],[22-24],[25-29] and show all the four bars for each country.

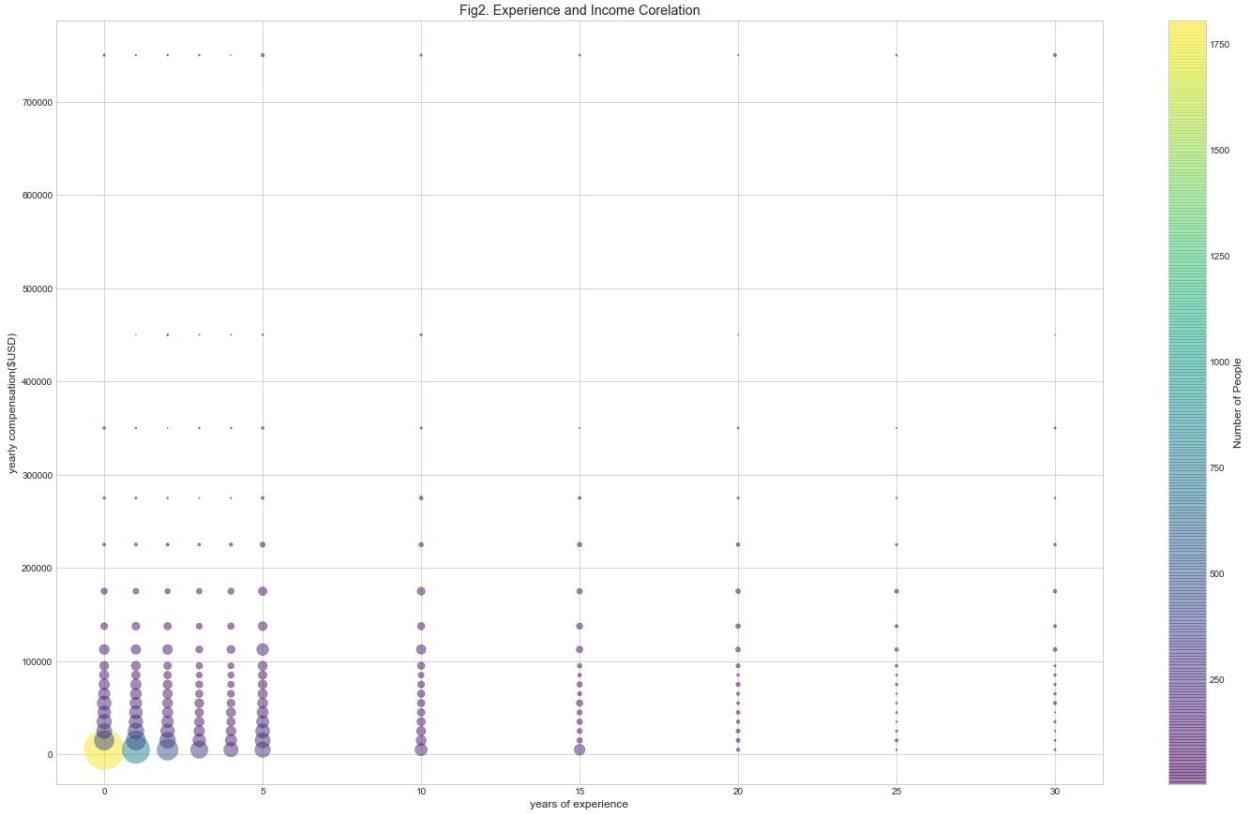
Fig1. Number of developers in each country



The plot is grouped bar chart which groups participants of all developers, 18-21, 22-24 and 25-29 for each country.

1. As for the number of participants of all developers(red color), USA, india and China rank top 3, and USA ranks first (while very close to india);
2. As for participants of 18-21(blue color) which are the youngest developers, India almost dominates, and USA and China rank second and third;
3. As for participants of 22-24(light blue color), India, USA and China rank top 3;
4. As for participants of 25-29(grey color), USA surpasses India and becomes NO.1, China still ranks third;
5. For USA participants, among 18-21, 22-24, and 25-29, more senior participants take part in this survey, while for India participants, more younger participants take part in this survey;
6. For participants of all countries, generally speaking, there are more senior participants(25-29) than younger participants(18-21).

To study the correlation between experience and income of people I produce the following scatter plot in which the circles show each group of people and the size of the circle shows the number of participants in that groupe.



1. For developers with few years of experience, the majority of yearly compensation is less than $100,000;
2. For those with more years of experience, the layout of yearly compensation is more balanced, which means that there are similar number of those people working with different yearly compensation
3. The percentage of high yearly compensation for more years of experience is higher compared with developers with few years of experience;
4. Most developers have experience less than 10 years, so, we may assume that this industry is a sunrise industry which has not been popular for a long time.

To seek into majors of participants I made a chart shows the population of major of the participants. Also used the chart to see how many users use AWS, GCP and Azure among all other cloud computing services. So reproduce the pie chart that shows the usage percentage of these three, among all other.

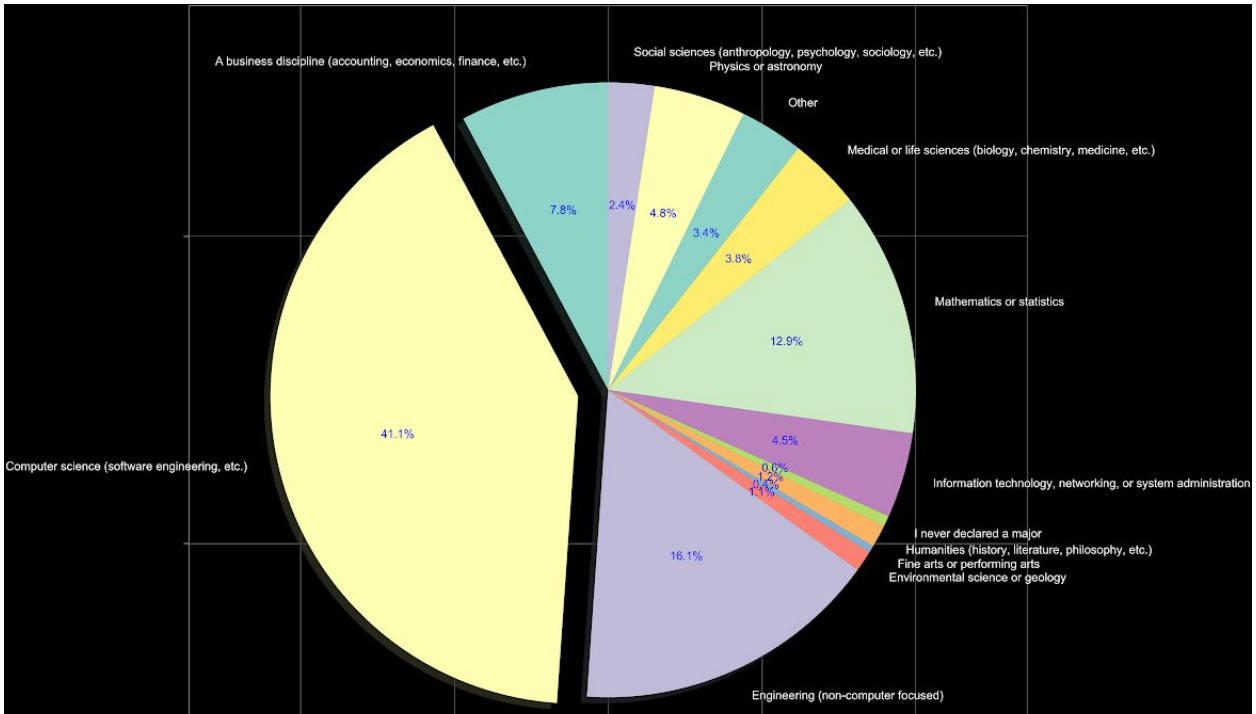


Fig3. Undergraduate Major of Participants

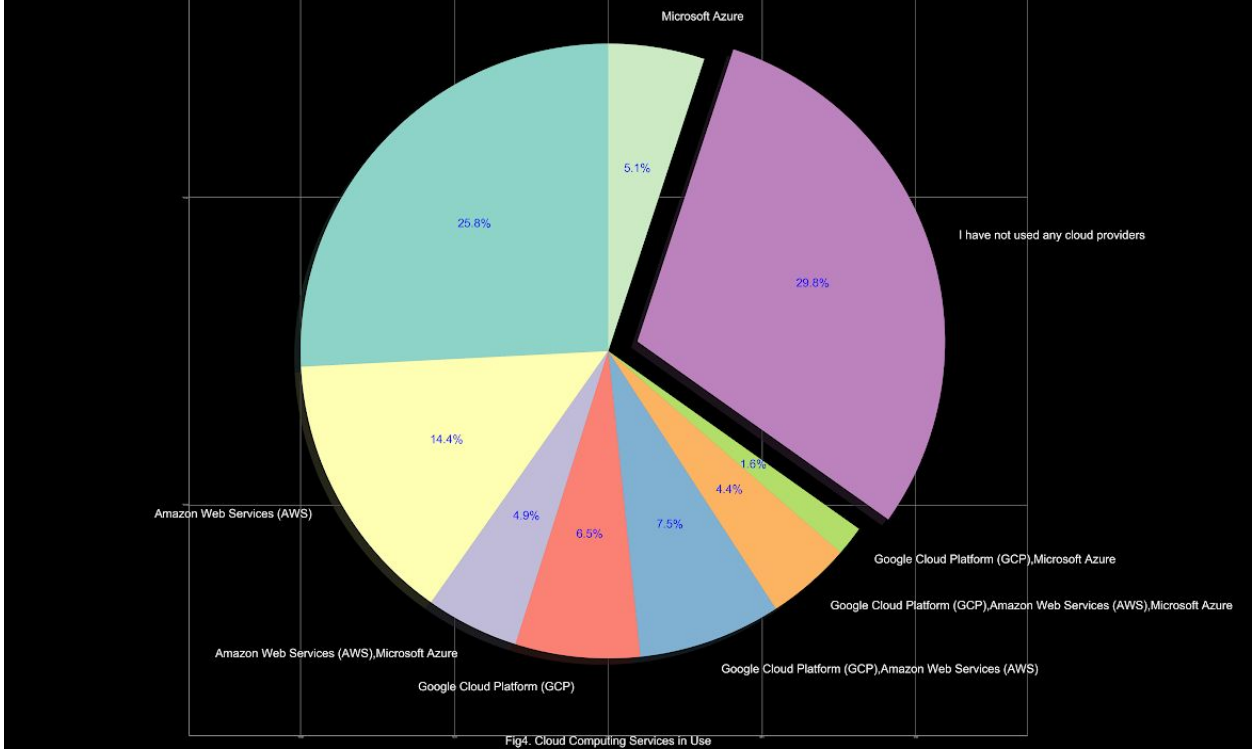


Fig4. Cloud Computing Services in Use

Analyze Fig 3,4 and give more insight about the dataset.

FIG3

1. Computer science(software engineering, etc.) dominates, which is indubitable;
2. Computer science(software engineering, etc.), Engineering(non-computer focused) and Mathematics or statistics rank top 3, which may prove that those knowledges of these three majors benefit the data science and machine learning career after graduation;
3. The major range is very broad, and even physics and humanities join here, which may indicate that data science and machine learning are very popular and interdisciplinary.

FIG4

1. About half of participants haven't used any cloud providers(29.8%) or provided related information(25.8%);
2. Among participants who have used cloud providers, Amazon Web Services are most popular, and AWS users(including those may also use other providers) account for 31.2%(14.4%+4.9%+7.5%+4.4%);
3. For combination usage, it's more likely that Google Cloud Platform and Amazon Web Services are used together, while there are very few cases that GCP and Microsoft Azure are used together.